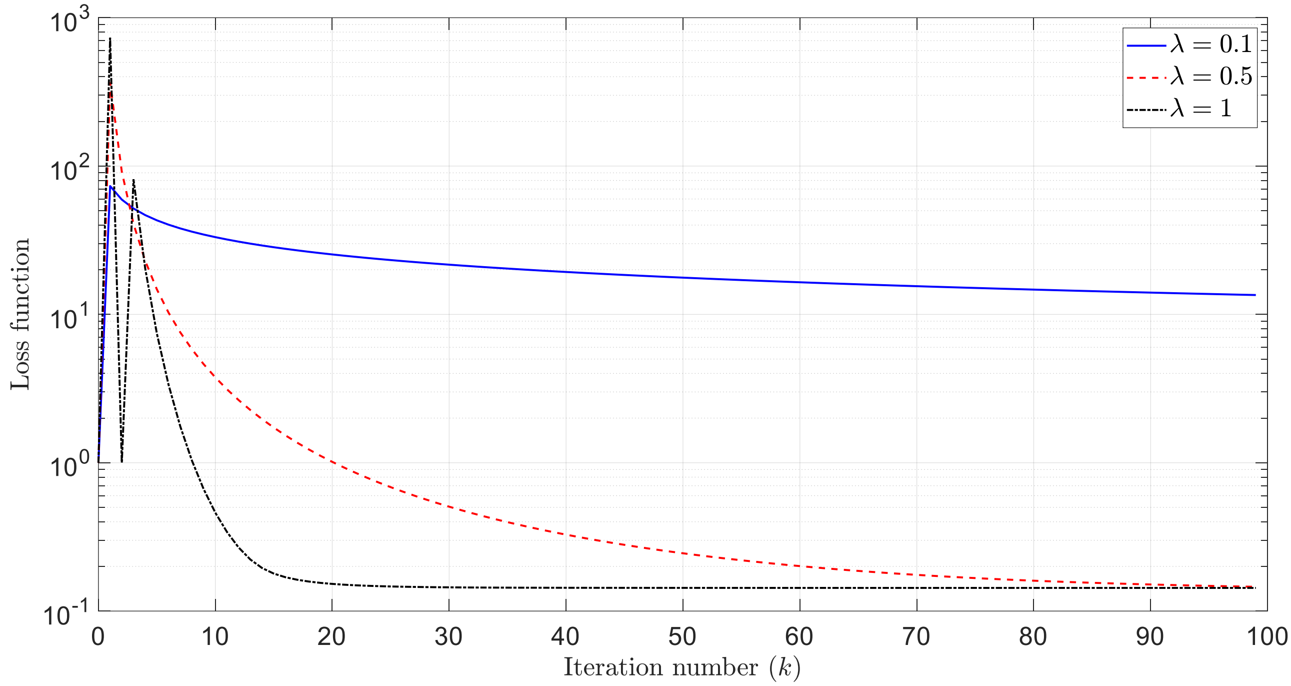
Computer Assignment 5

a) Decentralized gradient descent with 10 workers:

Following figure shows the convergence of the loss function with iteration number k. Figure characterizes the convergence for different values of regularization parameter λ.



b) Two-star topology with communication graph (1,2,3,4)-5-6-(7,8,9,10):

Convergence of Decentralized subgradient method and ADMM:

- We consider the following doubly stochastic matrix A with elements A_{ij} that is compatible with the given communication graph.

$$A_{ij} = \begin{cases} \dfrac{1}{\max\{d_i, d_j\}} & ; \quad if \;\; j \epsilon N_i \setminus \{i\} \\ 1 - \displaystyle\sum_{k=1,\;\; k \neq i}^{10} A_{ik} & ; \quad if \;\; i = j \\ 0 & ; \quad otherwise \end{cases},$$

where d_i is the degree of node i. For the decentralized subgradient method we used the following algorithm.

Decentralized Subgradient Algorithm:

At each iteration k, each user follows the following steps.

Step 1:

$$\bar{\boldsymbol{w}}_i^{(k)} = \sum_{j \epsilon N_i} A_{ij} \boldsymbol{w}_j^{(k)}$$

Step 2:

$$\boldsymbol{w}_i^{(k+1)} = \bar{\boldsymbol{w}}_i^{(k)} - \alpha^{(k)} \nabla f_i(\bar{\boldsymbol{w}}_i^{(k)})$$

For ADMM we used the following algorithm given in the reference (Makhdoumi & Ozdaglar, 2017).

ADMM algorithm for the given two-star communication graph:

At each iteration k, each user i follows the following steps with initializations $\boldsymbol{w}_i^{(0)}$ and $\boldsymbol{p}_i^{(0)} = \boldsymbol{0}.$

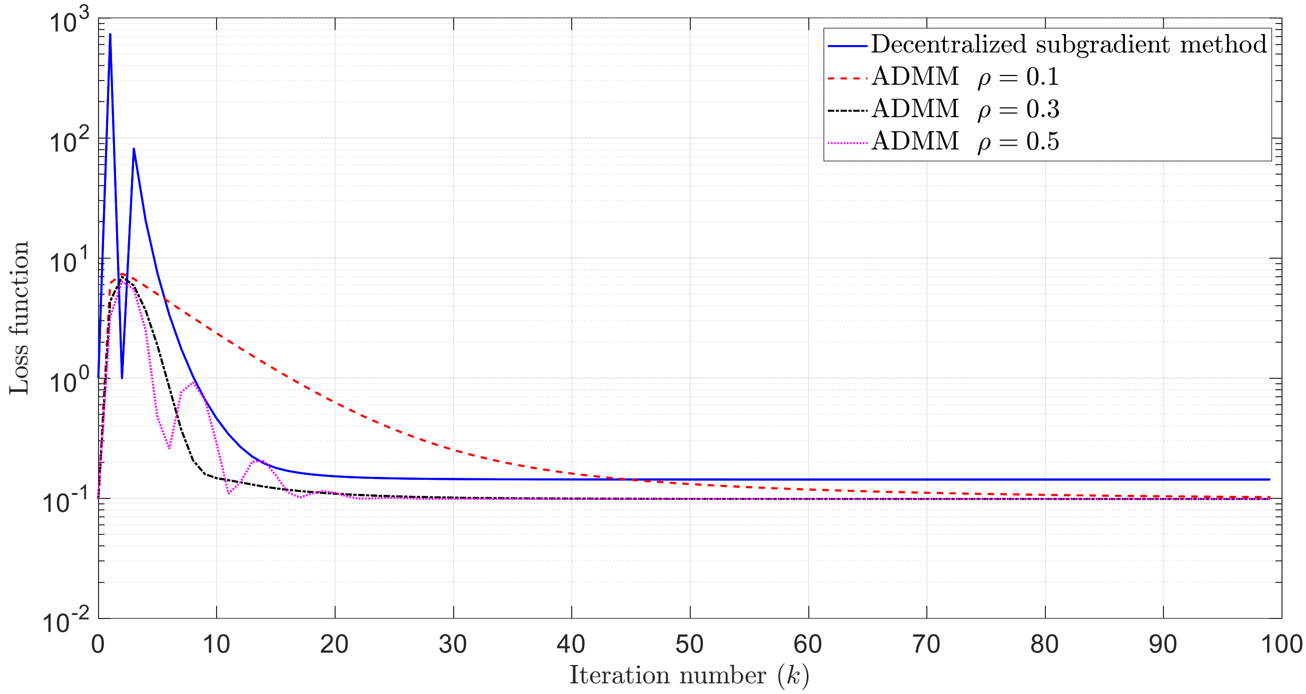
Step 1:

$$\boldsymbol{m}_i^{(k)} = \frac{1}{d_i + 1} \sum_{j \epsilon N_i} A_{ij} \boldsymbol{w}_j^{(k)}$$

Step 2:

$$\boldsymbol{w}_i^{(k+1)} = \operatorname*{argmin}_{\boldsymbol{w}_i} f_i(\boldsymbol{w}_i) + \sum_{j \epsilon N_i} \left(A_{ij} \boldsymbol{p}_j^{(k)T} \boldsymbol{w}_i + \frac{\rho}{2} \left\| \boldsymbol{m}_j^{(k)} + A_{ij} (\boldsymbol{w}_i - \boldsymbol{w}_i^{(k)}) \right\|_2^2 \right)$$

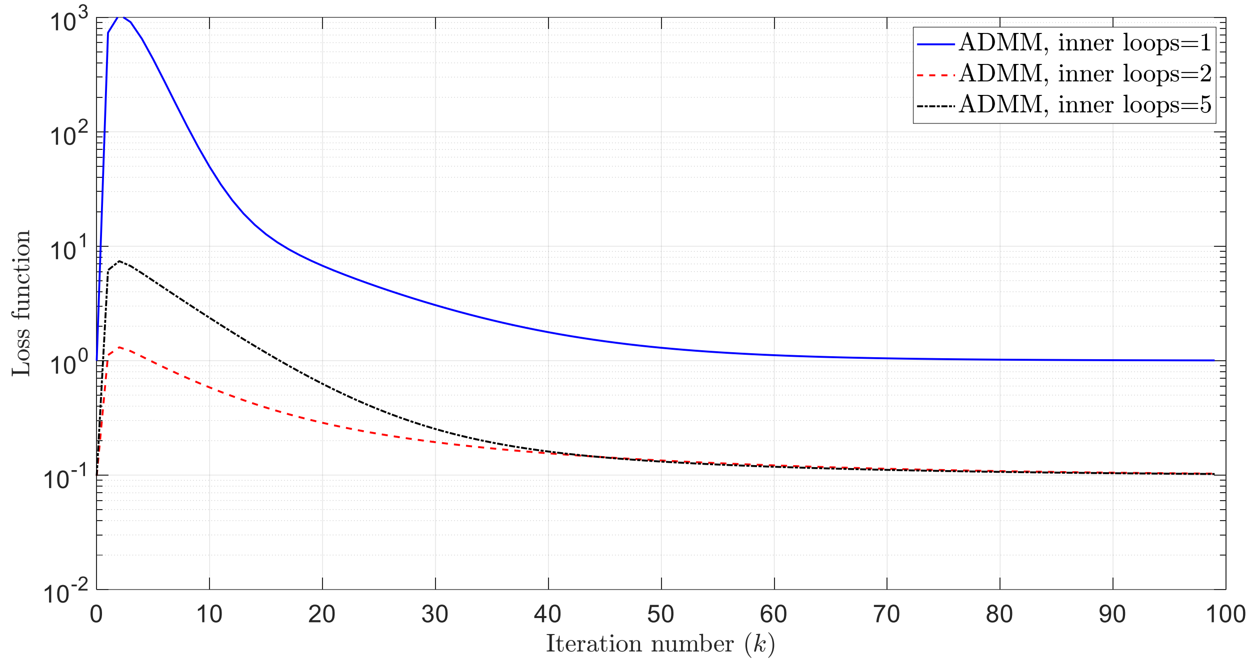
Step 3: $\boldsymbol{p}_i^{(k+1)} = \boldsymbol{p}_i^{(k)} + \rho \boldsymbol{m}_i^{(k)}$



Above figure shows that ADMM provides a better accuracy than the decentralized subgradient methods (however, according to the graph it is not a significant gain). Moreover, the rate of convergence increases when we increase the value of the penalty parameter ρ.

c) An approach to reduce *T* (number of iterations) with a marginal impact on the convergence

In step 2 of the ADMM algorithm, each user needs to run several gradient steps (inner loops) to minimize their own augmented Lagrangian functions. Instead of running large number of gradient steps (until the convergence of the inner loops) we propose to run few gradient steps. The resulting graph is given below.



References

Makhdoumi, A., & Ozdaglar, A. (2017). Convergence rate of distributed ADMM over networks. *IEEE Transactions on Automatic Control*.

Computer Assignment 6

Part a-)

In this problem we want to see affect of quantization and compression on communication efficiency. We used the code for CA5 for this question and repeated part a and part b as asked in the question definition. For Q1, we keep only a constant amount of elements and set others to 0. For Q2, we quantize in usual manner, taking less bits.

We see that using both Q1 and Q2 test error didn't end up far from train error. We used lambda = 1 for this part.

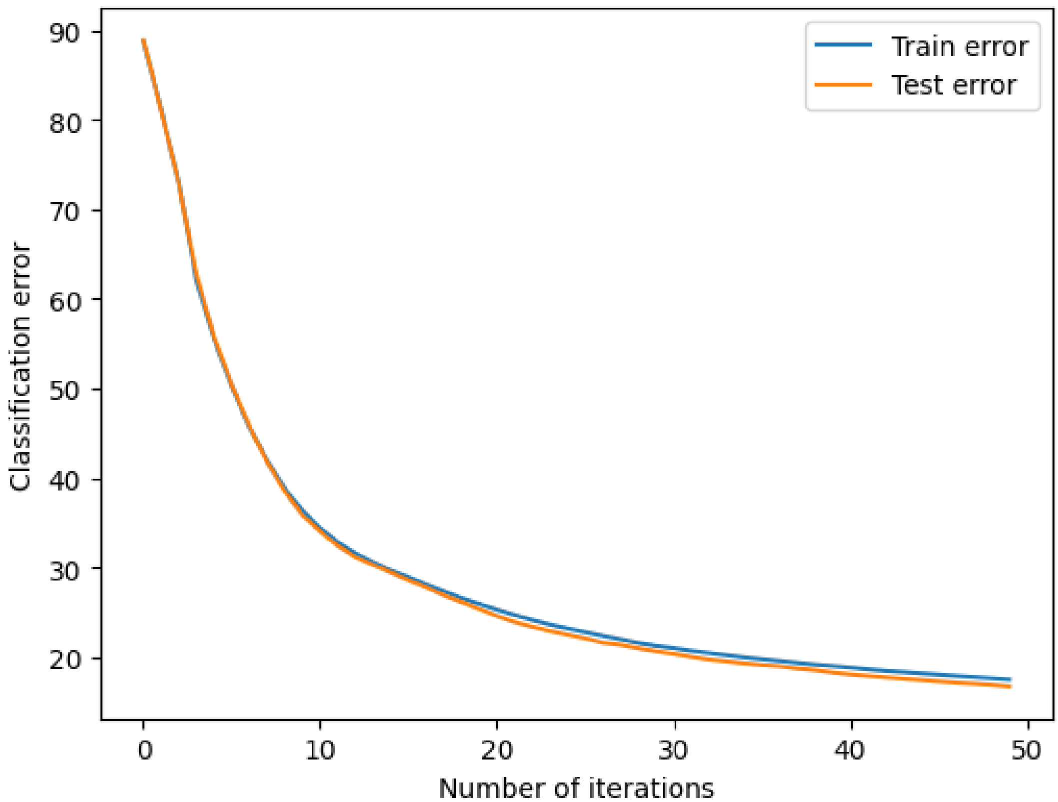


Figure 3. Comparison of Test and Train errors for Q1

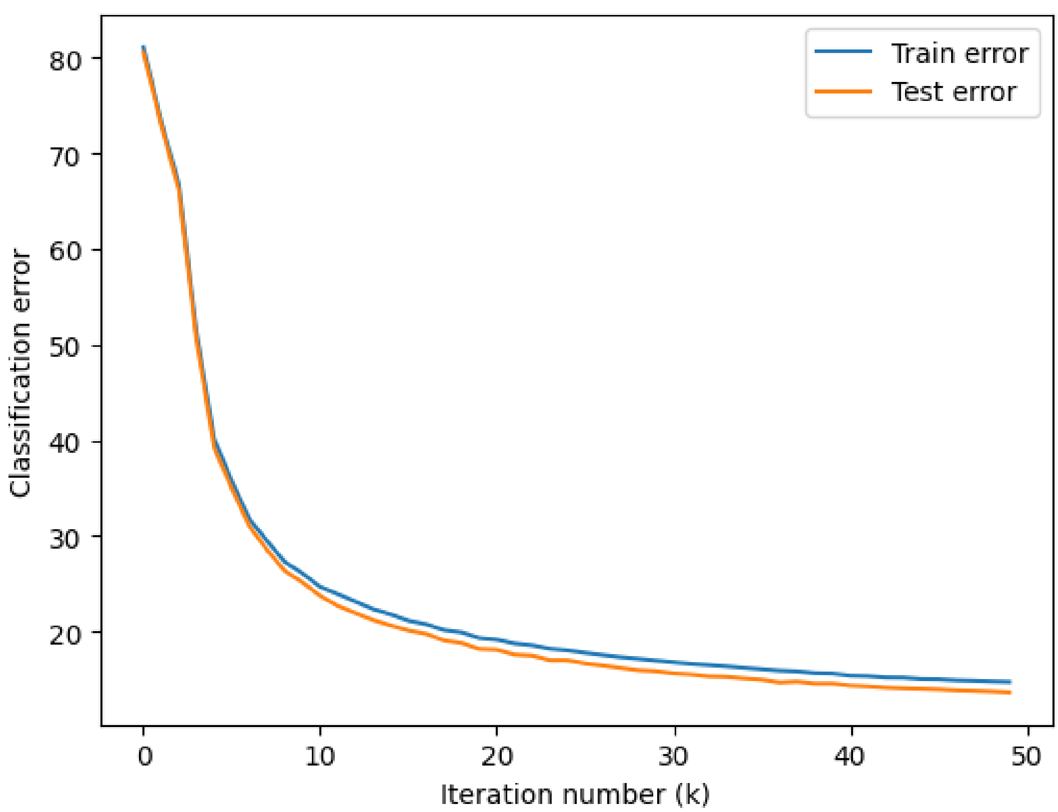


Figure 4. Comparison of Test and Train errors for Q2

From both figures we see that quantization doesn't affect us too badly. Therefore, it can be used to improve communication efficiency. Next, we will mention some other methods that can be used for communication efficiency.

Part b-)

In order to make SVRG and SAG communication efficient for large scale ML, we can use a decentralized model. A sparse graph would give a communication efficient method. Moreover, one can decrease amount of communication or take gradients locally. Doing local computations will result in decrease of communication.